Training Effects in Audio- Visual Integration of Sine Wave Speech

A Senior Honors Thesis

Presented in Partial Fulfillment of the Requirements for graduation *with distinction* in Speech and Hearing Science in the undergraduate colleges of
The Ohio State University

By

Megan Exner

The Ohio State University
June 2008

Project Advisor: Dr. Janet M. Weisenberger, Department of Speech and Hearing Science

ABSTRACT

Speech perception is a bimodal process that involves both auditory and visual inputs. The auditory signal typically provides enough information for speech perception; however, when the auditory signal is compromised, such as when listening in a noisy environment or due to a hearing loss, people rely on visual cues to aid in understanding speech. Visual cues have been shown to significantly improve speech perception when the auditory signal is degraded in some way. The McGurk and MacDonald study (1976) strongly supported the fact that speech is not a purely auditory process and that there is a visual influence even with perfect auditory input.

There is a growing interest in the benefit that listeners receive from audio-visual integration when the auditory signal is compromised. Remez et al, (1981) studied intelligibility when the speech waveform is reduced to three sine waves that represent the first three formants of the original signal. Remez discovered that sine wave speech is still highly intelligible even though a considerable amount of information was removed from the speech signal. Grant and Seitz (1998) looked at audio-visual integration performance of hearing impaired listeners by comparing a variety of audio-visual integration tasks using nonsense syllables and sentences. The study's results showed that even when the auditory signal is poor, speech perception is highly improved with the aid of visual cues. However, a large degree of variability was seen in the benefit that listeners receive from audio-visual integration. Further analysis suggested that at least some of this variability can be attributed to individual differences in listeners' abilities to integrate auditory and visual speech information.

Studies in our lab have explored the differences in benefit that listeners receive from visual cues during audio-visual integration. We propose that one source of the variability in the

benefit that listeners receive may be the overall amount of information available in the auditory signal.

A previous study in our laboratory, Tamosiunas (2007) explored the audio-visual benefit that listeners received using highly-degraded sine wave speech. Results of this study indicated that listeners received little benefit from the addition of visual cues and in some cases these cues actually inhibited speech perception. A possible explanation for the difficulties in speech perception found in this study was the degree of exposure subjects had to sine wave speech.

The present study explored whether the lack of audio-visual integration and benefit seen in Tamosiunas' (2007) study was a result of unfamiliarity with sine wave speech or whether this degree of auditory signal degrading inhibits audio-visual integration. To accomplish this, listeners in the present study were provided auditory and audio-visual training in sine wave speech perception. Results show that with training and exposure, speech perception performance did increase in both auditory and audio-visual conditions.

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Janet M. Weisenberger, for providing me with the opportunity to work with her. Through her support, guidance, and experience I was able to grow academically, professionally, and personally. I would also like to thank Natalie Feleppelle for her time and assistance throughout the year. Lastly, I would like to thank my friends and family for their support through this process.

This project was supported by an ASC Undergraduate Scholarship and an SBS Undergraduate Research Scholarship.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW

It has long been known that speech perception is a bimodal process that involves both auditory and visual inputs. The auditory input usually is sufficient for speech perception, but when the auditory signal is compromised, such as in a noisy environment or in the presence of a hearing loss, people rely on visual information to aid in understanding speech. Visual cues have been found to significantly improve speech intelligibility when the auditory information is degraded in some way. However, McGurk and MacDonald found, in a 1976 study, that we use visual information even when the auditory signal is not impaired.

McGurk and MacDonald's study was conducted by dubbing auditory syllables, such as the bilabial [ba], onto a video of a person saying a different syllable, such as the velar [ga]. By pairing conflicting visual stimuli with auditory stimuli, they were able to examine how participants integrated the audio and visual inputs together. In the auditory plus visual condition, a majority of subjects responded hearing [da], which is a fusion of the places of articulation for the visual syllable and the auditory syllable. This is referred to as a "fusion" response. "Combination" responses were also reported, such as [baga], in response to auditory [ga] paired with visual [ba]. These are responses that are a composite of the syllables in each modality. They are often formed because the strong visual information, such of that of a bilabial, cannot be ignored. When the visual information was taken away, in the auditory only condition, auditory syllables were heard correctly. The McGurk and MacDonald study strongly supported the fact that speech is not a purely auditory process and that there is a visual influence even with perfect auditory input. Today it is generally agreed that audio-visual integration occurs continually and

automatically, and it is important to evaluate both the audio and visual cues that can be obtained from the speech signal.

Auditory Cues for Speech Perception

The auditory signal provides a large amount of information that alone is adequate for intelligibility. Included in this is information on place of articulation, manner of articulation, and voicing for consonants. Place of articulation refers to where in the vocal tract the sound is produced, or the location of a constriction in the oral cavity. The manner of articulation describes how the sound is formed, or how the articulators make contact with one another. Voicing refers to whether the vocal folds are vibrating or not. If the folds are in a state of vibration during the sound production, then the sound is said to be voiced. If there is no vibration, then the sound is voiceless. All of this information is included in the spectral and temporal envelopes of the speech waveform (Borden *et al*., 2002).

Work by Shannon and his colleagues with degraded speech waveforms provided evidence that the acoustic waveform contained more information than necessary to identify a sound. In Shannon's study the temporal envelopes of speech token recordings were preserved, while reducing the spectral information (Shannon *et al*., 1995). To add ambiguity to the signal, band-limited noise replaced the fine-structure spectral information. These auditory stimuli were similar to sounds presented through cochlear implants. It was found that identification improved as the number of noise-bands increased. But surprisingly, even with only three bands of modulated noise-bands high levels of speech recognition could still be reached. These findings, supported the idea that the auditory speech signal is highly redundant, containing much information beyond the minimum needed for identification, and even a small amount of auditory cues can be useful in speech recognition.

Remez and colleagues (1981) studied intelligibility when the speech waveform was reduced to three sine waves that represent the first three formants of the original signal. By reducing the waveform in this manner they were able to degrade the auditory stimulus without adding noise. However, the speech quality of the signal is perceived as unnatural. Participants reported that the signal resembled science fiction sounds, artificial speech, computer beeps, and whistles, among other things. Nonetheless, it was discovered that sine wave speech can still be highly intelligible even though a considerable amount of information was removed from the speech signal. However, higher intelligibility was observed in sentences than in isolated syllables (Remez *et al*., 1981).

Visual Cues for Speech Perception

Research has also focused on the visual component of audio-visual integration to identify the mechanisms and cues provided by the visual input. The visual signal carries significantly less information than the auditory signal for sound identification. Place of articulation can be perceived to a certain degree, but there is considerably less information on manner of articulation, and absolutely no information on voicing in the visual signal (Jackson, 1988).

Having reduced information causes ambiguity in visual speech perception; a group of words or sounds may look the same visually and thus be visually indistinguishable. These groups are referred to as visual phonemes, or visemes. Items in a viseme group have an identical place of articulation, but differ in terms of manner and voicing, such as /p,b,m/ (all bilabials). Viseme categories are much more prominent for places of articulation near the front of the mouth, such as bilabials, labiodentals, linguadentals, and alveolars, and much less prominent for places of articulation near the back of the mouth, such as velars and glottals. Since a large number of English phonemes are not easily visible, speech-reading with no auditory input is

extremely difficult and varies with different talkers. It was found that some talkers were easier to speech-read, and thus created more viseme categories, whereas other talkers were more difficult to speech-read and created fewer viseme categories (Jackson, 1988). Other visual cues that may aid in speech perception by providing information are visible cues displayed by the individual talker, including eye, mouth, head movements, and gestures. These cues provide information on interest, attention, and involvement from the talker and the listener.

Audio-Visual Integration

Grant and Seitz (1998) looked at audio-visual integration of hearing impaired persons by comparing a variety of audio-visual integration measures in nonsense syllables and sentences. They used the Pre-Labeling Model of Integration (PRE) and the Fuzzy Logical Model of Perception (FLMP) as it fits to their data to determine whether individuals integrate auditory and visual cues with varying degrees of efficiency. Grant and Seitz found that variance observed across individuals in audio-visual speech recognition could be explained by differences in individual integration efficiency. The study's results also showed that even when the auditory signal is poor, speech is intelligible with the aid of a visual signal.

Theories of Audio-Visual Integration

Researchers have introduced various models to describe integration between the two modalities for optimal speech perception, and these models differ mainly on when integration occurs. As explained by Grant and Seitz (1998), the Pre-Labeling Model of Integration (PRE) suggests that information from the auditory and visual sources is combined before identification. The Fuzzy Logical Model of Perception (FLMP) suggests that separate decisions are made for each modality, then a final identification results from a mediating process.

The Pre-Labeling Model of Integration (PRE)

The PRE theorizes that audio-visual recognition is predicted by the ideal combination of information from auditory-only performance and visual-only performance. The model uses preserved unimodal information for the prediction in the multimodal case with no interference across modalities. The PRE can estimate integration efficiency and can be helpful in developing rehabilitative programs to maximize individuals' audio-visual speech recognition. Integration efficiency is measured by how far a person's performance deviates from a predicted audio-visual recognition score; the closer the performance is to the predicted score, the more optimally the individual is integrating. Audio-visual integration refers to the process of combining the information taken from the auditory and visual sources. This model believes that the efficiency of audio-visual integration is independent of a person's ability to extract auditory and visual information from speech. Integration occurs early, according to the PRE, prior to actual phoneme identification (Grant, 2002).

The Fuzzy Logic Model of Perception (FLMP)

The FLMP suggests that auditory, visual and audio-visual information are all independently assessed by the listener to create abstracts of the incoming information. These abstracts are then compared to summaries already in the memory in order to determine whether the cues match previous information already learned. These descriptions are then all integrated together and alternatives are formed. From these alternatives a decision and response is made based on the amount of support for each alternative. According to FLMP, integration occurs very late, after preliminary identification of the auditory and visual inputs (Grant, 2002).

Recent Audio-Visual Perception Studies

Several recent studies in our laboratory have investigated audio-visual perception of reduced-information speech-stimuli. Huffman (2007) and Andrews (2007) both looked at degraded audio stimuli that had been degraded in a manner similar to that of Shannon et al (1995). Both found good levels of identification performance. Tamosuinas (2007) looked at sine wave speech, similar to Remez el al (1981). Poor performance was seen even for the 3-formant (F1+F2+F3) condition. A possible explanation for this is that sine wave speech in citation style syllables just contains so little information that it is virtually unintelligible without linguistic content. However, Tamosuinas' participants had no prior exposure to sine wave speech. Anecdotal evidence in our laboratory shows that with increasing exposure, performance with these stimuli also increases. The present study addressed this by training subjects in recognizing sine wave speech stimuli. Subjects were tested pre-training in both audio-only and auditory-plus-visual conditions, and later tested post-training, to determine whether intelligibility increased after expose to the sine-wave stimuli, and whether audio-visual integration was facilitated.

CHAPTER 2: METHOD

Participants

Participants in the present study included ten listeners. Nine females and one male, ages 18-24, participated. All ten passed a hearing screening, and all also reported having normal or corrected vision. Three listeners reported some previous exposure to phonetics as part of their academic coursework. Participants were compensated $40 for their time.

Three talkers were used in this study, consisting of one male and two females between the ages of 20 and 23. All three reported being native speaker of midwestern American English speakers. Talkers were not compensated for their time.

Stimuli Selection

A limited set of eight syllables were presented as stimuli for the study. All syllables satisfied the following conditions:

1. The pairs of stimuli were syllables were minimal pairs; they differed only in the initial consonant.
2. All stimuli contained the vowel /ae/, used since it does not involve lip rounding or lip extension.
3. Multiple stimuli were used in each category of articulation, including: place (bilabial, alveolar, velar), manner (stop, fricative, nasal), and voicing (voiced, unvoiced).
4. All were presented citation-style, without a carrier phrase.

Stimuli

For each of the conditions the same set of single-syllable stimuli was used:

Bilabial: bat, mat, pat

Alveolor: sat, tat, zat

Velar: cat, gat

Visual and Audio Digital Video Recording

Digital recordings of three talkers presenting the stimulus multiple times were converted into degraded speech samples using a PRAAT version 4.4.29 script created by Chris Darwin, University of Sussex. This program produces sine wave speech by reducing the stimuli into three sine waves that represent the first three formants of the original signal. This degrading is achieved without adding noise to the signal.

Video Explosion Deluxe was used to dub the degraded auditory stimuli onto the corresponding visual stimuli of the talker saying the word. DVDs were burned, and consisted of 60 consecutive stimuli presentations, created from randomized stimulus lists to minimize memorization by participants. Twelve DVDs were created for each of the three talkers.

Stimulus Presentation

For the presentation of auditory stimuli JBL Reference 510 headphones were used. A 20-inch video monitor was placed one meter from the listener, who was seated in a sound-attenuating booth for stimulus presentation.

Testing Procedure

Study testing was done at a lab of Pressey Hall, which houses The Ohio State University's Speech and Hearing Department. Participants were instructed by the examiner verbally to read over a set of instructions. The instructions explained that the participants would hear a stimulus ending in "at" and were given a closed-set of eight possibilities. They were told they would be tested under two conditions, auditory alone and auditory plus visual, and to verbally respond to what they perceived.

Participants were individually tested in a sound attenuating booth, with the door closed. From their chair positioned in the booth they could see a video monitor placed outside the booth in front of a glass window. Auditory stimuli were transmitted through headphones inside the booth. Through an intercom system the examiner could record and score the participant's responses. Listeners were tested pre-training with 60 stimuli from each of three talkers, under auditory (A) and audio-visual (AV) conditions. Listeners then were provided with two hours of training with stimuli, under A and AV conditions, with feedback provided. Listeners then were tested post-training with 60 stimuli from each of three talkers, under A and AV conditions. The order of presentation of talkers and conditions (A, AV) randomized for all participants. Total training and testing for each participant took approximately five hours and was broken up into two sessions. To minimize fatigue, breaks were encouraged during sessions.

CHAPTER 3: RESULTS AND DISCUSSION

Results of the pre-tests and post tests were analyzed to determine whether training affected identification performance. Analysis also dealt with whether participants were trained for auditory speech perception or for integration ability.

Percent Correct Performance

Figure 1 shows the percent correct performance of listeners for the Auditory (A) Pre-test and Post-test and the Audio-visual (AV) Pre-test and Post-test. Performance is shown for each of the different talkers used in this study. Results were averaged for all listeners. There are several things worth noting in this figure. Results show that there is an improvement in the performance of listeners in the AV condition following sine wave speech training for all talkers. Improvement following training was seen for talkers LG and MO in both the A and AV conditions. However, improvement was variable across talkers and conditions. Variation across talkers supports the idea that individual talker characteristics can impact intelligibility of reduced-information speech signals (e.g., Andrews, 2007).

Figure 2 shows the percent correct performance of listeners for the A Pre-test and Post-test and the AV Pre-test and Post-test. Performance is shown for each of the different listeners used in this study, averaged across talkers. All listeners performed better in the auditory plus visual condition. All listeners, except one (BC), improved their performance from the pre-test to the post-test in the auditory only condition. And all listeners, except listener BD, improved their performance in the auditory plus visual condition between pre and post testing. It can be seen that individuals improved to varying degrees. Results show there was an improvement in

performance for both conditions following training for all listeners, but improvement was variable across listeners and conditions.

Improvement

Figure 3 shows the percent improvement (change in performance from pre-test to post-test) of listeners following sine wave speech training in the A and AV conditions for each talker. Results show that talkers LG and MO both yielded similar levels of improvement in both the A and AV conditions, while talker PV yielded significantly less improvement in the A condition.

Figure 4 shows the percent improvement of individual listeners following sine wave speech training in the A and AV conditions. Results show substantial variability across listeners in improvement in the A and AV conditions. Interestingly, those listeners showing a greater improvement following training in the A condition were not always those listeners showing improvement following training in the AV condition. This finding supports the argument that integration is a skill independent of auditory or visual alone performance (e.g., Grant & Seitz, 1998). If A-only performance were strongly linked to AV performance, one would expect improvements in both conditions following training.

Integration

Figure 5 shows the amount of AV integration facilitated by talker. Integration can be assessed by determining the degree to which auditory plus visual performance was better than auditory only performance. Results show that the amount of AV integration did not significantly improve following sine wave speech training. Again, this finding underscores the idea of independence of integration ability.

Statistical Analysis.

Statistical analysis was run using a 2-factor, repeated measures ANOVA, to reveal any significant findings in the study. A significant main effect of presentation condition was found, $F(1, 9) = 245.342$, $p<.001$, with AV performance better than A. A significant main effect of training was also found, $F(1, 9) = 17.214$, $p=0.002$, suggesting a facilitating effect of training. No significant interaction was found, $F(1, 9) = 2.570$, $p=0.143$.

T-tests (Bonferroni- corrected) were also run. Results indicate that training did not optimize audio-visual integration, $t(9) = -0.967$, $p=0.359$, such that the amount of benefit between A and AV did not change from pre-test to post-test. Thus, although training produced significant improvement for audio-only conditions, $t(9) = -3.525$, $p=0.006$, and for audio-plus-visual conditions, $t(9) = -3.555$, $p=0.006$, the amount of benefit between A and AV was not increased by training.

Comparison with Previous Studies

The findings of this study are in contrast to those of Tamosiunas (2007). Tamosuinas found that integration was hindered in some cases by adding auditory input to the visual signal, resulting in higher performance in visual-only conditions than audio-visual conditions. Therefore, his study suggests that there may not be enough information in sine-wave speech to facilitate audio-visual integration. It is important to note the levels of performance at the start of this study (from the pre-training test) compared to the levels of performance Tamosuinas found. Participants in this study performed at comparable levels to Tamosuinas' participants in auditory-only conditions. However, Tamosuinas' participants performed at much lower levels in audio-visual conditions than subjects in the present study, even prior to training. This

discrepancy could be attributable to individual differences across listeners in the two studies, such as a greater amount of noise in Tamosuinas' stimuli. It is also possible that the difference is due to differences in stimuli between the two studies. Tamosuinas used four different sinewave configurations (F0, F1, F2, and F0+F1+F2). The single-sine wave stimuli contained much smaller amounts of information than the three-sine wave stimuli, and therefore are considered less intelligible. It is possible that subjects in Tamosuinas' study became so frustrated with the highly unintelligible single sinewave stimuli that they decided that the entire task was impossible.

CHAPTER 4: SUMMARY AND CONCLUSION

Results of testing indicated significant a main effect of training, with overall post-training performance substantially better. Specific comparisons indicated significant improvement in both auditory alone performance and in auditory plus visual performance. However, the amount of integration did not show a significant change as a function of training, suggesting that training effects were centered on auditory intelligibility. An explanation for this is that a different type of training regimen may be necessary to tap into integration skills. Interestingly, the degree of benefit from training varied substantially across listeners. Some listeners improved tremendously, a few had relatively little improvement, and one even got worse after training. This supports the findings of Grant and Seitz (1998) regarding individual differences in integration ability.

Listeners in this study were trained for a relatively small amount of time, approximately two hours. However, improvements still were found. It is possible that a longer training period would result in even greater improvements. Therefore, future studies with longer training periods should be performed to determine whether an extended training period could facilitate both auditory intelligibility and AV integration skills.

Overall, the results of this study suggest that training with highly degraded auditory stimuli can lead to substantial improvements in intelligibility even when large amounts of information have been removed. Results of this study have clinical implications for the design of aural rehabilitation programs for hearing impaired persons. Training those with hearing loss may help an individual make the most use of his or her residual hearing. Since audio-visual integration is thought to be a skill independent of auditory-alone performance, then perhaps, training for integration should be separate and different from training for auditory listening skills.

Support for this idea would impact the amount of training and services available to those with hearing impairments, and therefore, may result in more use of any residual hearing.

CHAPTER 5: REFERENCES

Andrews, B. (2007). *Auditory and visual information facilitating speech integration.* Senior Honors Thesis, The Ohio State University.

Borden, G.J., Harris, K.S., & Raphael, L.J. (2002). Speech Science Primer: Physiology, Acoustics, and Perception of Speech [4th ed]. Baltimore, MD: Lippincott Williams & Wilkins.

Grant, K.W. & Seitz, P.F. (2002). Measures of auditory-visual integration for understanding: A theoretical perspective (L). *The Journal of the Acoustical Society of America, 112 (1),* 30-33.

Grant, K.W. & Seitz, P.F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *The Journal of the Acoustical Society of America, 104 (4),* 2438-2450.

Huffman, C. (2007). *The role of auditory information in audiovisual speech integration.* Senior Honors Thesis, The Ohio State University.

Jackson, P.L. (1988). The theoretical minimal unit for visual speech perception: Visemes and coarticulation. *The Volta Review, 90 (5)*, 99-114.

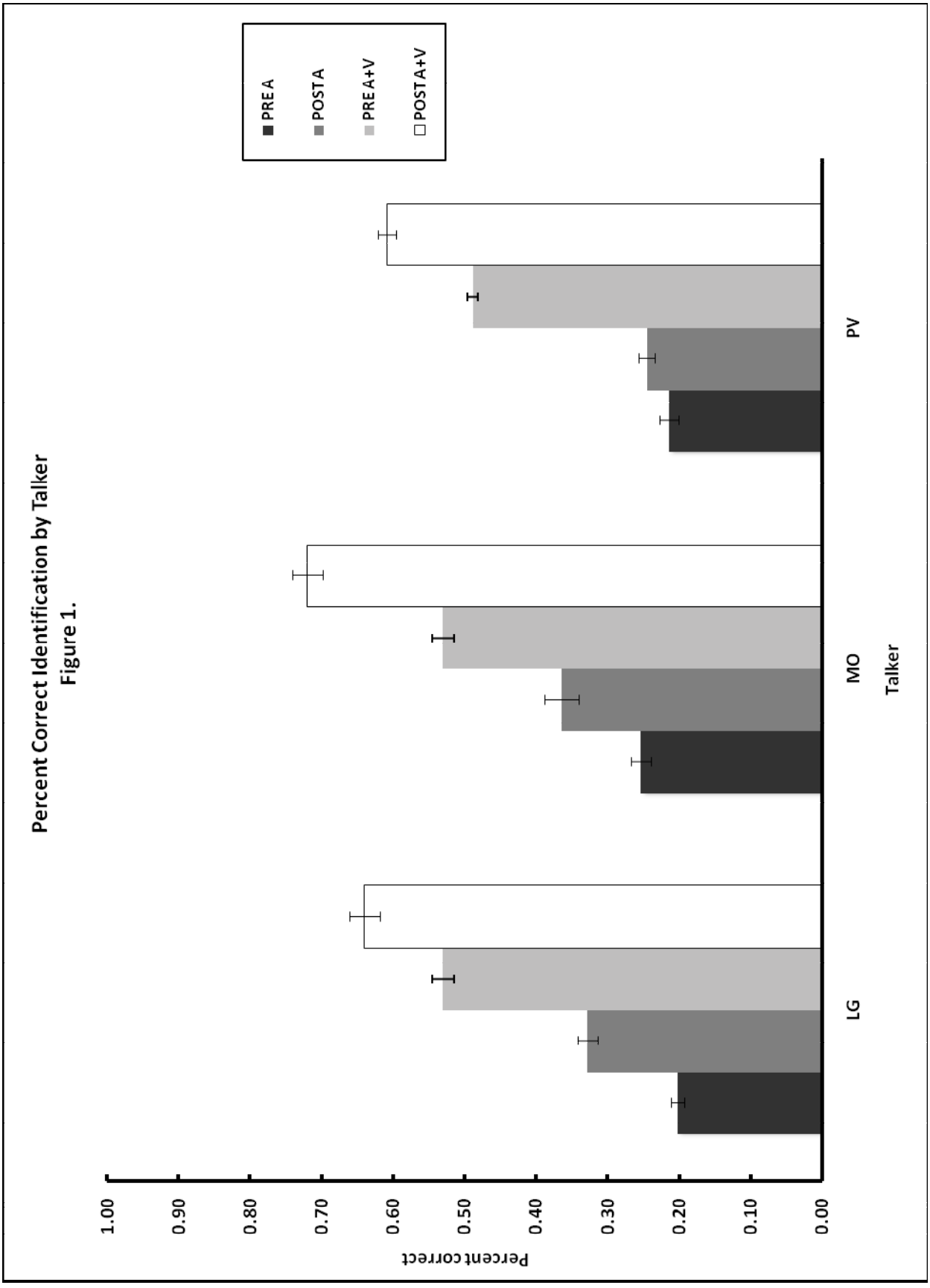
McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746-748.

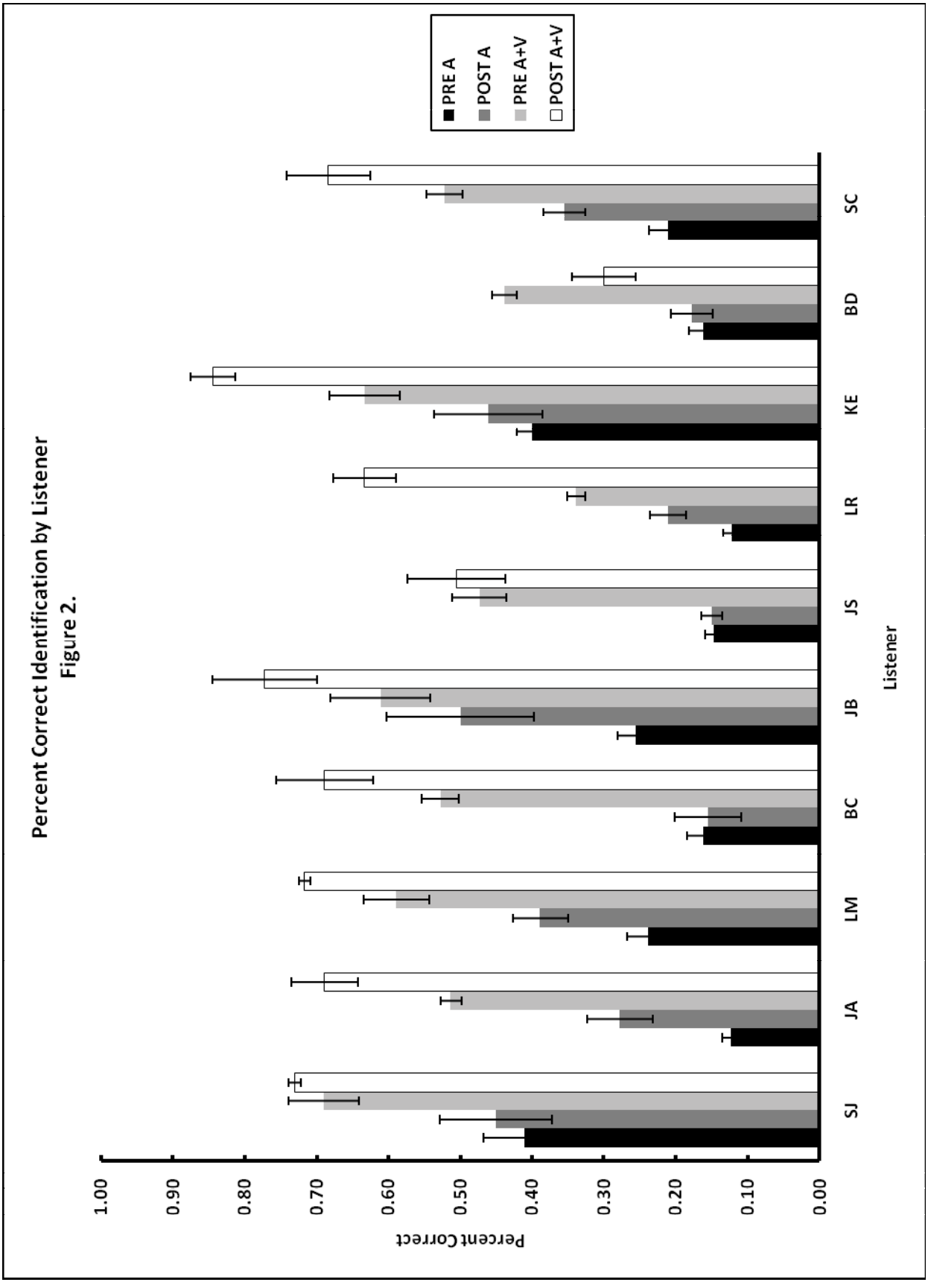
Remez, R.E., Rubin, P.E., Pisoni, D.B., & Carrell, T.D. (1981). Speech perception without traditional speech cues. *Science, 212 (4497)*, 947-950.

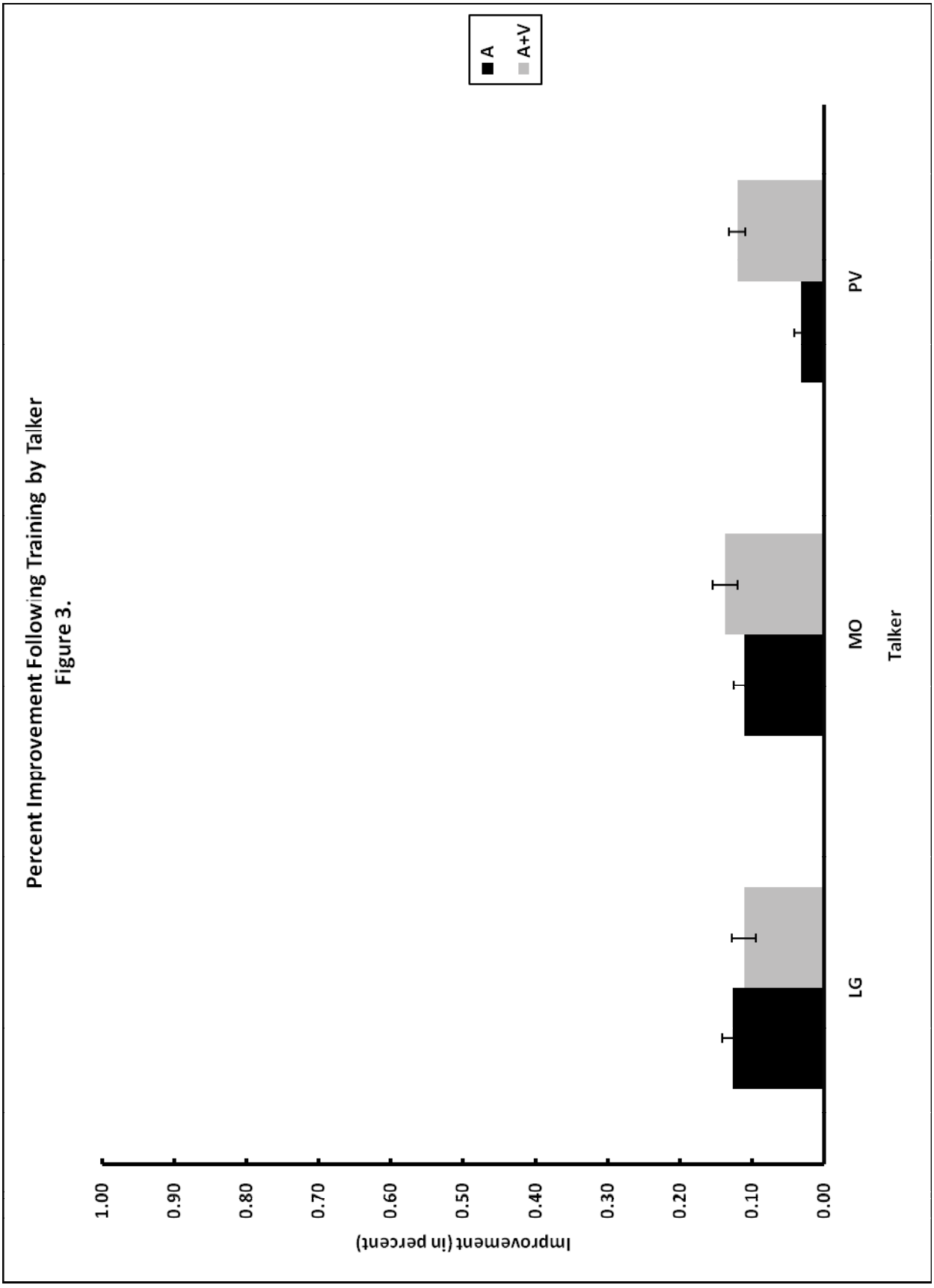
Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270,* 303-304

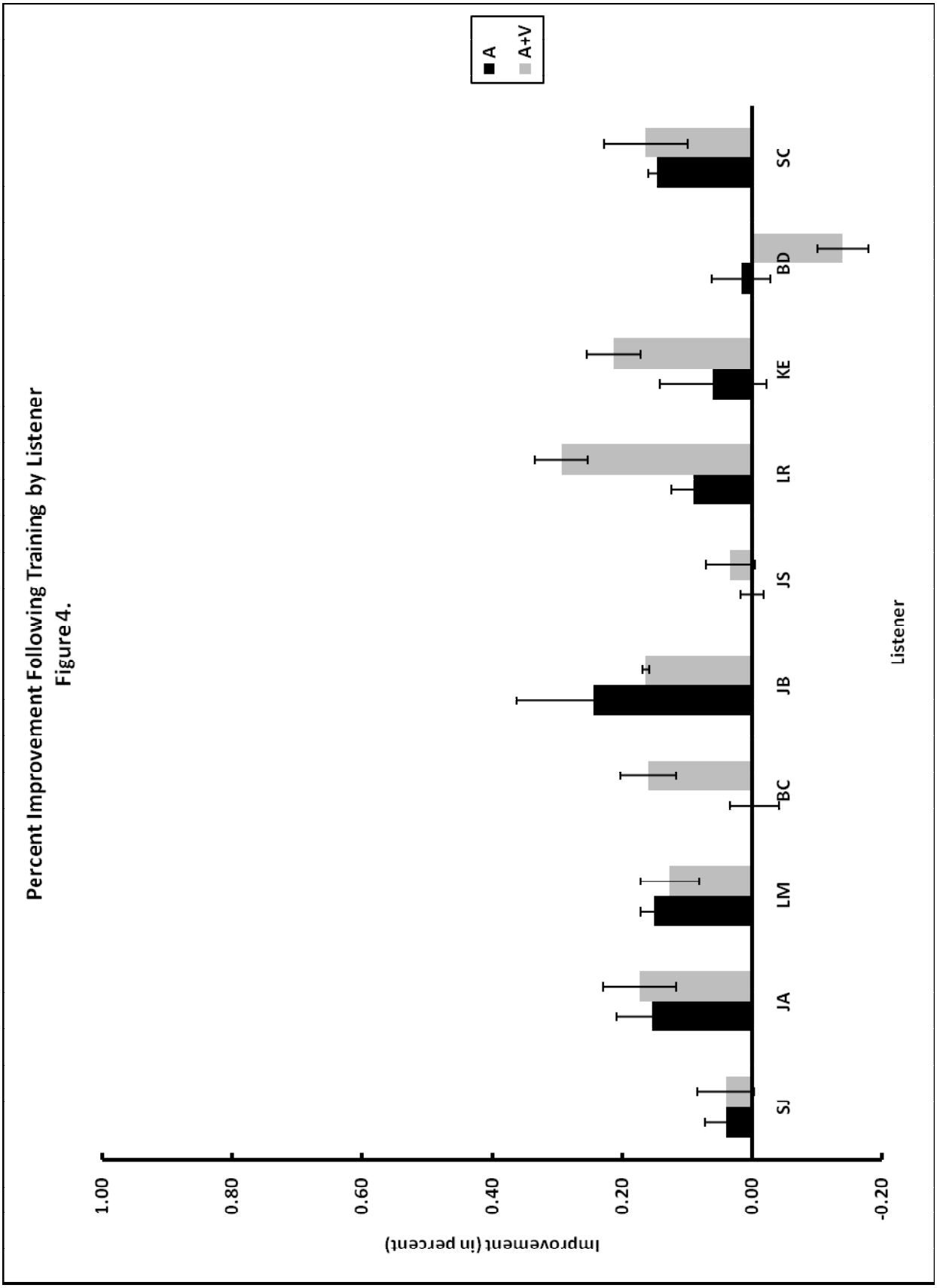
Tamosuinas, M. (2007). *Auditory-visual integration of sine-wave speech.* Senior Honors Thesis, The Ohio State University.

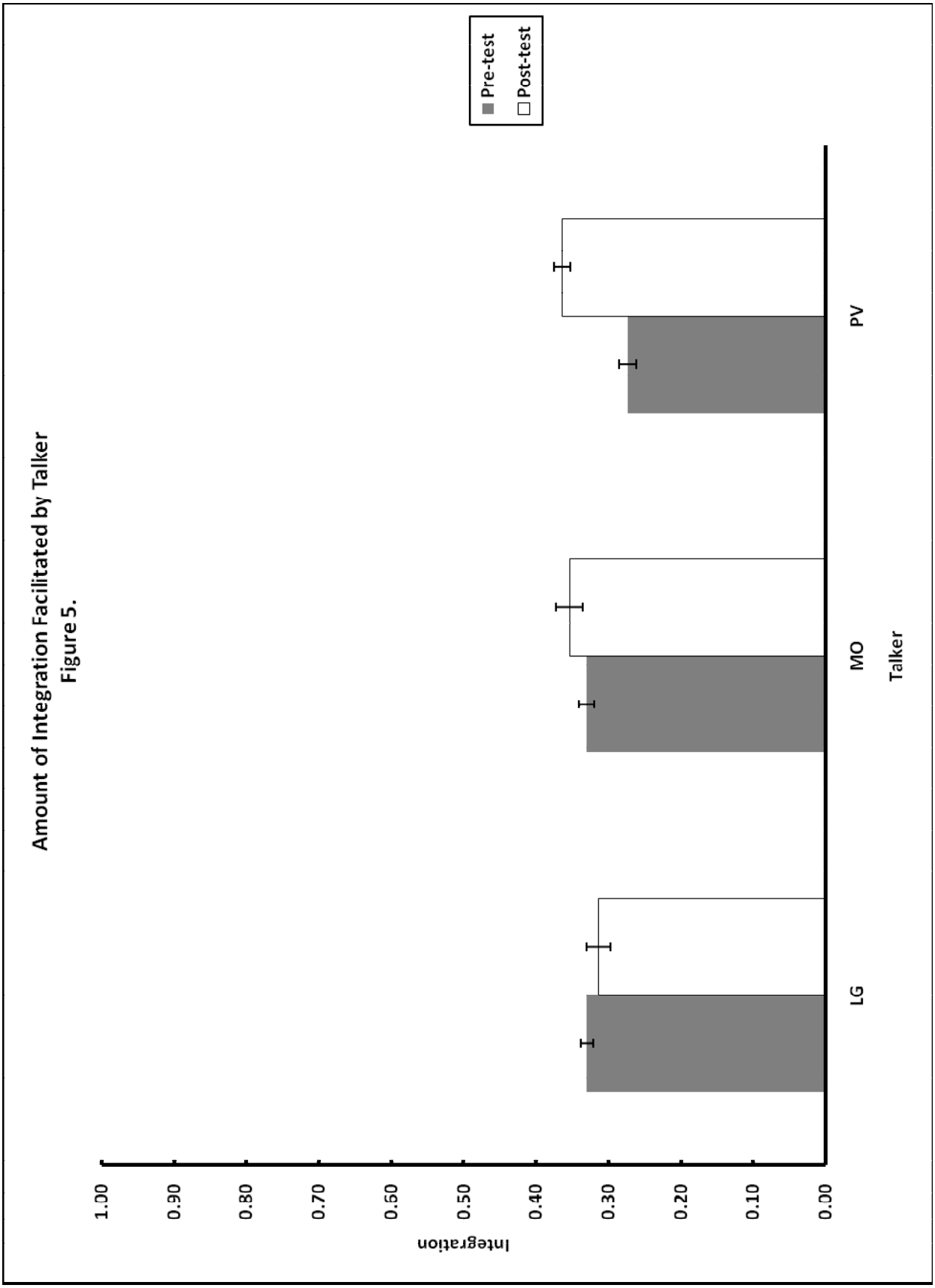
LIST OF FIGURES


Percent Correct Identification by Talker
Figure 1.
PRE A
POST A
PRE A+V
POST A+V
Percent correct
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
LG
MO
PV
Talker


Percent Correct Identification by Listener
Figure 2.
PRE A
POST A
PRE A+V
POST A+V
Percent Correct
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
SJ
JA
LM
BC
JB
JS
LR
KE
BD
SC
Listener


Percent Improvement Following Training by Talker
Figure 3.
A
A+V
Improvement (in percent)
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
LG
MO
PV
Talker


Percent Improvement Following Training by Listener
Figure 4.
A
A+V
1.00
0.80
0.60
0.40
0.20
0.00
-0.20
Improvement (in percent)
SJ
JA
LM
BC
JB
JS
LR
KE
BD
SC
Listener


Amount of Integration Facilitated by Talker
Figure 5.
Pre-test
Post-test
Integration
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
LG
MO
PV
Talker